Ad Hoc Retrieval with Harris SENTINEL

Margaret M. Knepper, Gregory J. Cusick, Kevin L. Fox, Ophir Frieder, Robert A. Killam

Harris Corporation, Information Systems Division
P.O.Box 98000, MS W3-7755
Melbourne, FL 32902
(407) 984-6443

mknepper@harris.com
kfox@harris.com

1.0 INTRODUCTION

Harris Information Systems Division (HISD) focuses on information retrieval support for various Government agencies. In our customers' applications, efficient (in terms of processing time) retrieval rates are critical, and highly accurate but relatively few documents retrieved is the norm. Our SENTINEL approach addresses our customers' needs.

This is HISD first participation in the Text REtreival Conference (TREC). Our team participated in the category C (large data set) manual Ad Hoc track of the Sixth Text REtrieval Conference (TREC-6). Throughout TREC-6, we made modifications to enhance the performance of our system. We improved both the processing time and document retrieval. This paper is an overview our efforts for TREC-6.

2.0 SENTINEL OVERVIEW

2.1 Retrieval Components

SENTINEL is a fusion of multiple retrieval engines, integrating n-gram technology, a Vector Space Model, and a neural network.

2.1.1 N-Gram

SENTINEL employs a n-gram filter based on Julian Youcum's work with least-frequent trigraphs [1]. SENTINEL moves a n-character sliding window over a document while recording the frequency of occurrence of different n-character combinations. A general frequency table is built from a corpus of training documents, representative of the document collection. Relevant documents are rapidly identified by looking for the occurrence of the least-frequent n-gram of a search string in the document. SENTINEL used a 3-character sliding window for TREC.

2.1.2 Vector Space Model

SENTINEL also uses the Vector Space Model (VSM) to represent documents in a n-dimensional vector space. Words appearing in the document training corpus are represented as vectors in the n-dimensional vector space. A vector for each document is constructed based on the terms in a document. A query is considered to be like a document, so a document and query can be compared in the vector space. Documents whose content, as measured by the terms in the document, correspond most closely to the content of query are judged to be the most relevant [2]. The documents are retrieved through keyword, word clusters (series of words), and example document queries mapped into the n-dimensional vector space. The documents whose vectors are a minimal distance from the query's vector are retrieved.

SENTINEL implemented a new algorithm for our VSM component during TREC-6 which reduced the document representation processing time from three minutes to approximately five seconds per document.

2.1.3 Neural Network

A neural network is used within SENTINEL to train the word vectors in our VSM. The neural network is based on Kohonen's Self-Organizing Map neural network [3], [4]. It is an unsupervised learning algorithm that organizes a high-dimensional vector space based on features in the training data so that items with similar usage are clustered together. This clustering accounts for individual words appearing in close proximity in a document and hence an implied similar or related meaning.

2.2 Ad Hoc Queries

We employed SENTINEL in a multi-level processing approach to manually query for the Ad Hoc topics in the TREC data. A series of queries was created for each of the Ad Hoc topics. In less than two months, our three-person team was able to perform over 900 manual queries over the entire document corpus. Table 2.2-1 shows a summary of the query distribution. The average number of queries per topic was 19. The maximum number of queries for a topic was 90, and the least number of queries for a topic was 6.

For each Ad Hoc topic, keywords and phrases were input to the n-gram component of SENTINEL to obtain rapid information retrieval extraction. This high-level filter yielded initial document screening information. This technique was especially useful for phrases not

Table 2.2-1. Number of Queries

Pass	Number of Queries
n-gram	361
Vector Space Model Pass 1	196
Vector Space Model Pass 2	130
Vector Space Model Pass 3	218
Subset Queries	43
Total	948

represented in the training corpus, and hence the vocabulary of SENTINEL's VSM.

Keywords, word clusters, single documents, and document clusters were input to SENTINEL's VSM component as queries. Queries constructed by SENTINEL's VSM can be broadly or narrowly focused, depending on the keywords, phrases, and example documents used in the queries. SENTINEL's VSM was used for high accuracy query retrieval and document scoring. The score is obtained by computing the distance between the vectors representing the query and the document. Document scores ranged from -1 to +1. Negative scores indicated an irrelevant document. Scores for relevant documents[1] ranged from approximately .45 to 1. The closer to 1, the better the document matches the query.

Experiments have shown that SENTINEL's VSM strongest performance results from the use of example documents and document clusters. Initial document examples were obtained

1. Relevant documents are documents we felt met the topic criteria. Irrelevant documents may have been close to the topic, but in our opinion, did not meet the topic's requirements provided in the Description and Narrative accompanying each of the Ad Hoc topics.

from the n-gram filter and topic related articles found on the web. As the passes were completed, top query results were reviewed and identified as relevant or irrelevant. Relevant documents from the query were input to the next pass of SENTINEL's VSM.

New hardware also improved query performance. First, we obtained a SPARC Ultra which increased our RAM from 90 MB to 128 MB. Next, we obtained 16 GB of additional disk space. The new hardware increased the average VSM Pass 1 query processing from 13,000 documents/day to 57,000 documents/day.

2.2.1 Subset Queries

Instead of processing over the entire document corpus for each topic and query, we created subsets for several topics. Subsets were created by the n-gram filter for topics 302 (Polio), 304 (Endangered Species, Mammals), and 347 (Wildlife Extinction) because they contained keyword phrases. Queries were then run over the subsets.

2.3 Ranking

Query results for each topic were processed through multiple methods to enhance the ranking performance. The techniques included relevant document score enhancement and elimination of irrelevant documents and queries. Our algorithm was used to rank query results based on: the number of times the document was selected, highest score, lowest score, and average score.

We reviewed the query results throughout all the passes. As previously mentioned, document scores range from -1 to +1, and the closer to +1 the better the document. As we reviewed queries we set a lower limit for the acceptable documents to reduce the amount of processing required to determine the top 1000 documents. The acceptable lower limit increased as higher rated documents were found. Each topic had it's own acceptable lower limit.

2.3.1 Document Score Enhancement

Relevant documents were used as examples that were incorporated into the next query pass. Originally, the query document wasn't processed as a possible query match. We thought other queries would identify the example documents as good matches and raise the score. However, other queries failed to reintroduce the relevant example documents. The relevant example document was being penalized since it was not acknowledged as relevant and processed. We modified the algorithm to process the example query document as a possible match.

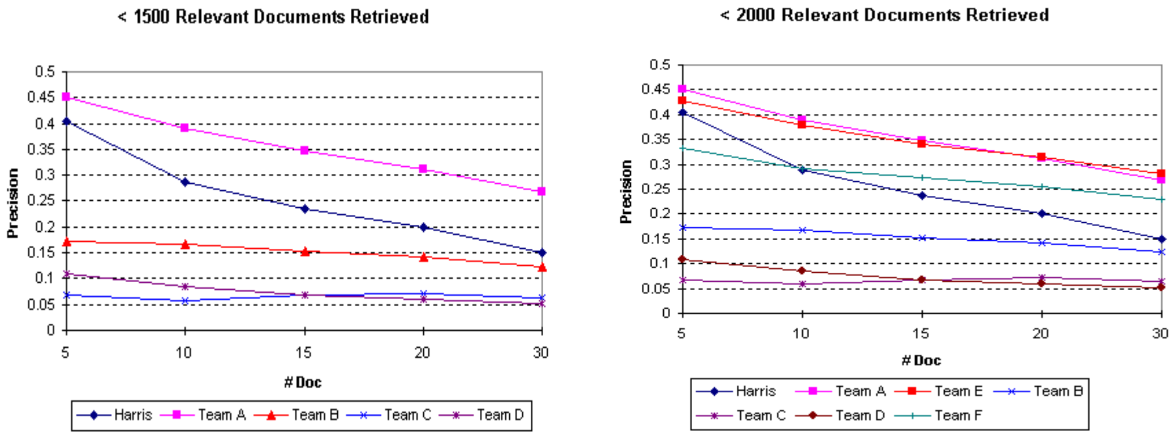
2.3.2 Document Elimination

Top documents were reviewed and irrelevant documents identified. Irrelevant documents were filtered from subsequent queries. Removal of the higher-scoring irrelevant documents allowed lower scoring documents to be accepted on the final result list.

A set of queries was constructed for each topic. For each topic a separate list of query results were maintained. We reviewed individual query results and results from combinations of queries. While reviewing results it became obvious that certain queries did a better job of retrieval then others. Leaving out an irrelevant query permitted more documents from better queries to be added to the final results.

3.0 TREC-6 RESULTS

SENTINEL was designed to yield efficient high precision for a small retrieval set. We compared our precision results with teams that retrieved a similar number of relevant documents. Figure 3.0-1 shows our system maintains a high level of precision for the top 30 documents retrieved. Our precision is higher than other teams that retrieved more relevant



Team	# Relevant Documents Retrieved
Harris	725
Team A	1,300
Team B	1,172
Team C	1,156
Team D	203
Team E	1,659
Team F	1,831

Figure 3.0-1. Harris maintains a high level of precision

documents.

Time constraints and interest-level limit the user to reviewing the top documents before the determining if the results of a query were accurate and satisfactory. The analyst needs the representative documents at the top of the list. SENTINEL retrieves the relevant documents high on the list. SENTINEL permits the user to build and tailor the query as he further defines the topic. The system permits movement from a generic search to a specific topic area through query inputs.

4.0 ANALYSIS

We reviewed several of our ideas to see if they were successful.

4.1 Irrelevant Document Removal

Reviewing the documents for topic relevance is a subjective task. Our analysis revealed that we had marked documents as irrelevant that had been judged for NIST as relevant. It quickly raised the question "Should we remove irrelevant documents?" The answer is "yes", but we will be more judicious about the documents we remove in the future. We re-ran several topics removing and not removing the documents we had previously judged as irrelevant. Removing irrelevant documents improved our precision score and also gave us additional relevant documents on the final list. But we were hurt in one topic in particular for removing too many "irrelevant" documents.

4.2 Pass Improvements

As expected, the results improved with successive passes. Reviews of each of the passes did reveal that several queries retrieved a high percentage of documents in the first pass. This allows the analyst to get good results with the first pass and build upon the query if desired. Table 4.2-1 shows some examples of the percentage of relevant documents retrieved by SENTINEL relative to the final number of relevant documents retrieved for the topic.

Table 4.2-1. Percentage Relevant Documents Retrieved by SENTINEL for Pass 1 and 2

Topic	% Relevant Documents Found 1st Pass	% Relevant Documents Found 2nd Pass	Total # Relevant Documents in TREC	Topic	% Relevant Documents Found 1st Pass	% Relevant Documents Found 2nd Pass	Total # Relevant Documents in TREC
301	73%	88%	474	308	100%	100%	4
303	25%	50%	10	315	31%	89%	67
304	15%	57%	226	324	89%	85%	162

4.3 Irrelevant Query Removal

We are able to select the combination of queries to represent the topic. The data was re-run using all the queries. There wasn't much difference between using all the queries and selected queries.

Individual query examination reveals which queries are retrieving relevant documents. Additionally, a relevant document may be a top document in one query because the scores are lower in the query. Some relevant documents were found reviewing the results from individual queries rather than the results from the entire collection of queries.

4.4 Time Management

We pondered the question - "Should we try and focus on a couple of topics?" or "Should we try to do our best on all the topics?" We decided not to focus on any particular topic and divided the topics among team members. Sometimes we spent too much time on individual topics due to our own interest! Some topics were neglected due to time and difficulty. We needed to accept the fact that not every topic will have hundreds of documents and don't try to find relevant documents that don't exist.

In some cases after we found a large number of relevant documents for a topic, we turned our attention to other topics. This was true for topics 301, 302, 324, and 330. The next question we pondered - "Should we spend more time on topics where we were having success?" We remained focused on our original goal of trying not to focus on specific topics.

In the cases of topics 303, 317, and 344, we couldn't accept the fact that we didn't find a lot of documents. However, in reviewing the dates of the articles in the corpus we began to realize that the data would be limited and quickly turned out attention to other topics.

4.5 Lost Documents

Most of our initial queries for a topic were generic, using keywords as initial query inputs. As we reviewed the documents retrieved by the queries, we interpreted which documents represented relevant topic documents. We proceeded to tailor and modify the query through the relevant documents input to the next SENTINEL pass. Our document selection controlled the query results. The documents started to represent a variety of clusters in our vector set. As we moved into more specific document query examples we found more document clustering, and consequently the document scores started to rise. This caused us to lose some of the generic lower-clustering and lower-scoring documents found in the early passes.

5.0 ANTICIPATED IMPROVEMENTS

5.1 Negative Queries

SENTINEL's VSM component contains a negative query feature. Example irrelevant documents are identified and the system automatically removes this type of document. Modifications to SENTINEL's VSM removal algorithm didn't give us time to fully test the feature.

Negative query removal was performed during the query. However, it needs to be performed during the ranking where the irrelevant documents are identified. We are currently modifying this feature.

5.2 Stronger Integration

Our process was intensely manual. We've already modified the ranking algorithm to perform integration of the results from the n-gram filter and VSM. It allows the user to weight the different retrieval methods based on the individual query, scales results (accounting for higher then average scores from some documents in the n-gram filter), and penalizes documents only found by one method.

5.3 Improved Interface

We are improving the interface to make relevant document classification and query building easier. SENTINEL will be accessed through a web browser.

5.4 Similarity Measure

Currently, SENTINEL's VSM calculates the query score as the distance between the vectors representing the query and the document. New methods for scoring will be explored.

5.5 Multiple Views

This year one person worked on a query for the entire TREC duration. Next year we plan on rotating the queries among team members. We feel this will give us different perspectives and ideas to further develop the queries.

5.6 3-D Viewing

3-D viewing of different document aspects is being explored using Harris VisualEyes tool. This tool enables us to view clustering and document location with respect to the query. This enables a user to locate additional documents relevant to the query, at least by the aspects being viewed.

Figure 5.6-1 shows an example of the 3-D viewer. We tested a query for retrieving stories about the McVeigh trial using CNN web documents. The query keywords were: McVeigh, trial, Oklahoma City, and bomb. Document locations are represented in space by a box, additionally in this view documents determined as relevant by SENTINEL display the document name next to the box. Clustering of documents can be observed in several areas:

- Cluster 1: Clustering of the McVeigh trial stories. Plus additional stories related to the topic, not identified by SENTINEL
- Cluster 2: Bosnia stories dealing with bombing are near the keyword "bomb".
- Cluster 3: O.J. Simpson trial stories appear near the word "trial".

6.0 SUMMARY

We learned a lot in our first TREC. We have a base retrieval system on which we can build and improve. Integration of the n-gram and VSM, an improved scoring algorithm, and TREC scoring experience gives us confidence for an improved TREC experience next year.

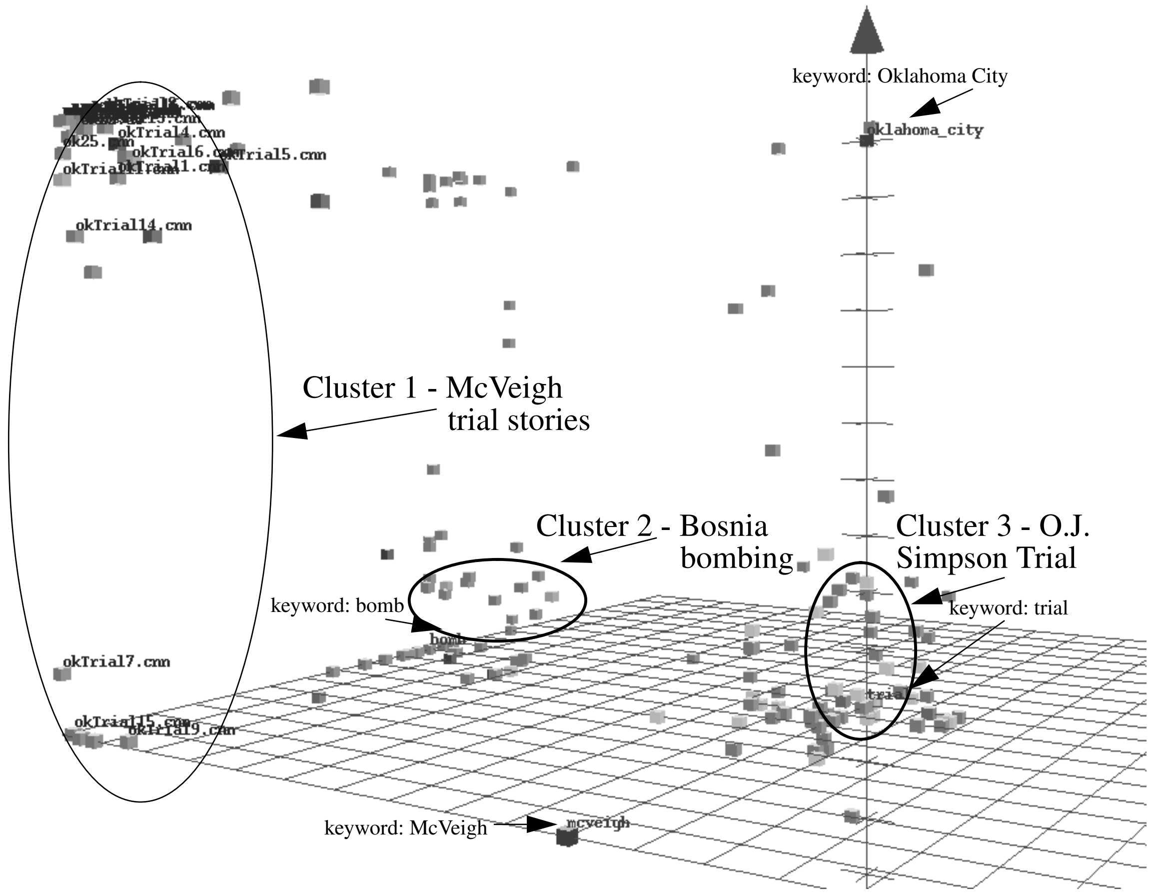


Figure 5.6-1. 3-D viewing of query to find a stories about the McVeigh trial

7.0 REFERENCES

[1] Julian A. Youcum, "High-Speed Text Scanning Algorithm utilizing Least-frequent Trigraphs." IEEE Symposium on New Directions in computing. 1985

[2] David A. Grossman and Ophir Frieder, Information Retrieval: Adhoc Query Processing, Kluwer Academic Publishers, Boston, to appear

[3] Teuvo Kohonen, Self-Organization and Associative Memory. Springer-Verlan, New York, 1984

[4] Simon Haykin, Neural Newtorks: A Comprehensive Foundation, IEEE Press, Macmillan College Publishing CO., New York, 1994